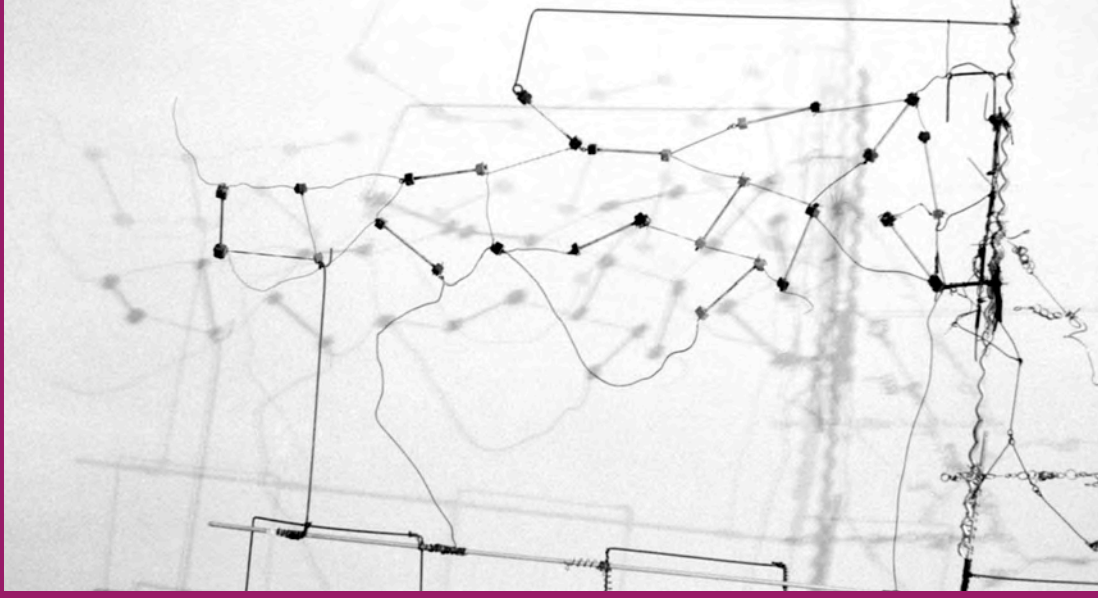
Territoires numériques



Marta Severo,
Alberto Romele

Traces numériques et territoires

MINES ParisTech
Presses des Mines

Marta Severo, Alberto Romele, *Traces numériques et territoires,* Paris, Presses des Mines, Territoires numériques, 2015.

© Presses des MINES - TRANSVALOR, 2015

60, boulevard Saint-Michel - 75272 Paris Cedex 06 - France

presses@mines-paristech.fr

www.pressesdesmines.com

ISBN : 978-2-35671-206-6

© Couverture : Emiliano Neri

Dépôt légal : 2015

Achevé d'imprimer en 2015 (Paris)

Tous droits de reproduction, de traduction, d'adaptation et d'exécution réservés pour tous les pays.

Traces numériques et territoires

Collection Territoires numériques

Dans la même collectin

Devenirs urbains
Maryse Carmes, Jean-Max Noyer

Les débats du numérique
Maryse Carmes, Jean-Max Noyer

Publié avec le soutien de l'EA GERiiCO, du CNRS et du CIST

Traces numériques et territoires

Marta Severo, Alberto Romele

MINES ParisTech

Introduction

Marta Severo, Alberto Romele

Ces dernières années, les nouvelles technologies ont profondément changé les territoires. Ce qui rend ce changement particulièrement intéressant est le fait qu'il affecte à la fois les territoires dans leurs matérialités et la façon de les étudier et de les gérer. Les médias numériques sont intéressants dans la mesure où toute interaction qui les traverse laisse des traces qui peuvent être enregistrées, analysées et visualisées. Cette traçabilité intrinsèque promet, si contrôlée par une méthodologie adéquate, de fournir une source nouvelle de données pour l'étude des territoires. Face à l'abondance de ces nouveaux types de données, plusieurs études empiriques ont été réalisées, mais une réflexion théorique sur l'emploi de ces données dans les études territoriales est encore faible. Cet ouvrage vise alors à combler au moins partiellement ce vide théorique en proposant des pistes d'analyse de la rencontre entre trace et territoire dans le cadre d'approches disciplinaires différentes.

D'abord, il est apparu nécessaire d'approfondir la question des méthodes et notamment de nouvelles méthodes digitales développées pour traiter ce type de données. Pour cette raison, l'ouvrage débute par un texte de Richard Rogers qui s'interroge sur l'impact des *big data* sur le travail du chercheur en sciences humaines et sociales. Rogers veut aller au-delà des critiques épistémologiques et éthiques de ces données pour poser l'accent sur les problèmes liés à leur traitement. Dans la promesse d'un regard distant sur les phénomènes sociaux, les *big data* risquent plutôt, selon l'auteur, de produire une forme de cécité et de mettre au second plan les questions liées à la signification de ces données. Rogers revient, enfin, sur la question de la méthode adaptée aux traces numériques et insiste sur la différence entre natifs numériques et numérisés qui peut s'appliquer tant aux données qu'aux méthodes dans les études en SHS.

Plusieurs chercheurs en STS (études des sciences et technologies), informatique et sociologie politique ont attiré également l'attention sur la convergence entre méthodes sociologiques et méthodes digitales. Dans leur contribution, Marres et Gerlitz s'intéressent plutôt aux similitudes et aux différences entre les outils grand public en ligne, qui s'offrent à l'analyse de données, et les techniques de recherche sociologique. Avec le terme «méthode interface», les auteures veulent indiquer précisément des méthodes émergentes que les chercheurs en sciences sociales ne peuvent pas dire être leurs, mais qui rentrent suffisamment en résonance avec les démarches qui leur sont familières.

Dans leur intervention, Severo et Romele passent de la question de méthodes à la question des données et introduisent l'étiquette de «*soft data*». Les *soft data* peuvent être définies comme des données disponibles sur Internet, facilement accessibles et récoltables. Elles sont constituées principalement par les nouveaux types de données issues du Web 2.0 (Facebook, Twitter, fils RSS, etc.) qui s'offrent au décideur public comme une source originale et riche

d'informations sur les phénomènes sociaux qui ont lieu dans un territoire. Par rapport au terme *big data*, parler de *soft data* permet de souligner que les données numériques peuvent être bien utiles pour les politiques publiques même si elles n'ont pas le volume, la vitesse et la variété des grandes bases de données. Par rapport aux *open data*, un regard orienté *soft* permet d'accueillir toutes les données qui ne sont pas libres de droit.

La première partie de l'ouvrage se termine avec le chapitre de Jos de Mul qui se concentre sur le passage des données numériques à l'identité numérique. Dans une perspective philosophique, de Mul s'interroge sur l'identité humaine à l'époque des grandes bases de données informatisées. Si les traditions ont toujours joué un rôle important dans la configuration de nos identités, c'est notre rapport aux traditions qui a changé. Dans les sociétés prémodernes, les traditions allaient de soi ; dans les sociétés modernes, les populations sont conscientes de la diversité historique et culturelle des traditions, et habiter une tradition est question de choix existentiel. Enfin, dans les sociétés postmodernes, la tradition est devenue plus une commodité qu'un choix existentiel. Selon l'auteur, les TIC, et en particulier les bases de données informatisées, ont été à ce regard déterminantes. D'un point de vue technique, les bases de données se trouvent derrière chaque site Web ; d'un point de vue anthropologique, les bases de données deviennent de plus en plus des métaphores conceptuelles qui structurent notre expérience du monde et de nous-mêmes.

Dans la deuxième partie, cet ouvrage se concentre sur des questions plus théoriques soulevées par la rencontre des traces numériques et des territoires. D'abord, il est nécessaire d'approfondir la notion de territoire pour en saisir ses nouvelles dimensions liées au numérique. Dominique Boullier, dans son texte, propose « une description progressive de huit propriétés des territoires les plus classiques pour discuter à chaque fois les nouvelles dimensions qui peuvent être ouvertes ». Pour chacune des propriétés, il part d'un agencement topographique pour définir un agencement topologique puis un agencement chronologique généré par les traces des plates-formes numériques. L'auteur met l'accent sur le fait que ces agencements et leurs combinaisons complexes doivent être pris en compte pour la gestion et le gouvernement des territoires.

L'article de Boris Beaude aborde la question du rapport entre trace numérique et espace et y répond dès la première ligne de son chapitre en affirmant que « les traces numériques sont éminemment spatiales ». Dans son texte, l'auteur souligne le potentiel inédit des traces pour les sciences qui s'intéressent à la dimension spatiale du social, tout en posant également l'accent sur les nombreuses questions soulevées par leur emploi. Entre autres, Beaude analyse la question de la visibilité, les enjeux des traces entendues comme big data, l'opacité des processus et surtout le risque de réductionnisme et d'appauvrissement des analyses qui risquent de réduire le Monde en général, et l'individu en particulier, à ses traces. En conclusion, Boris Beaude, comme d'autres auteurs dans ce volume, pose l'accent sur l'opportunité offerte par ce type de données mais en même temps sur la nécessité de prendre en compte leurs limites.

Franck Cormerais porte son regard sur un objet spatial spécifique, la ville, et propose la notion d'hyperville – à ne pas confondre avec la *smart city* – comme une alternative possible à la «gouvernementalité algorithmique». L'hyperville serait un «système local où la relation aux données, c'est-à-dire aux traces-signes, opère une reconfiguration dans la redistribution spatiale et temporelle». Néanmoins, un tel concept n'est pas naïvement optimiste. Il s'agit plutôt d'une approche «pharmacologique» – poison et remède, selon le double signifié du mot grecque pharmakon – des données, qui offrent à la fois des opportunités et des dangers. Dans les différents domaines de l'hyperville – politique, sociale, artistique, industrielle, d'opinion, économique – existent en effet de bonnes et de mauvaises pratiques digitales. Dans le cas du social, par exemple, là où la décision publique concerne l'intégration, les mêmes données peuvent fonctionner pour la création de réseaux sociaux ou pour leur segmentation. La nature de l'hyperville, en somme, consiste en des données, traces-signes, qui peuvent être utilisées de deux manières diamétralement opposées. Il en revient finalement à nous de valoriser ses aspects les plus contributifs. Un renouveau politique de la ville peut en effet modifier nos conditions politiques d'existence.

Si l'intervention précédente assume une posture quasiment neutre à l'égard des données numériques, Carmes et Noyer proposent une vision décidemment plus critique par rapport à ce qu'ils appellent le «plissement numérique du monde». A son origine, l'omniprésence de capteurs de données. Comme conséquence, des nouvelles (bio)politiques et économiques du pouvoir. La ville représente à ce propos un cas paradigmatique : elle ne sera vraiment smart que lorsque ses habitants, ses interfaces et applications seront domestiqués, *i.e.* seront tous devenus des «capteurs infatigables […] des activités de la vie quotidienne». Les auteurs soulignent opportunément que la datafication du monde ne peut que passer par la création et l'alimentation d'un désir de données. Discours et rhétoriques, utopies et idéologies, jouent à ce propos un rôle fondamental. Carmes et Noyer s'appuient notamment sur les trans et les post humanismes, qui tentent à leur avis de fournir un horizon – une justification ou des conditions d'existence et efficacité, nous pourrions dire – anthropologique du plissement numérique du monde.

Enfin, dans une troisième et dernière partie, cet ouvrage se confronte aux conséquences de l'utilisation des traces numériques pour l'aménagement et la gestion des territoires. Aujourd'hui, le décideur public doit intégrer les données traditionnelles avec les nouvelles données générées, selon une approche *bottom-up*, par les acteurs du Web 2.0. On assiste ainsi à l'avènement d'un nouvel impératif participatif dans l'élaboration et la mise en œuvre des politiques territoriales.

Matthieu Noucher propose d'utiliser le concept de trace numérique pour renouveler la cartographie critique. Une telle approche est aujourd'hui en crise face à de nouveaux phénomènes et de nouveaux acteurs de l'information géographique, comme Google Maps et OpenStreetMap, qui ont changé radicalement la manière de produire des cartes. Selon l'auteur, dans cette optique de renouvellement, de nouveaux cadres méthodologiques doivent être mis en place pour analyser la fabrique cartographique et dans ces cadres, une entrée

par les traces numériques peut contribuer à mettre en avant la question de l'intentionnalité cartographique et notamment du rapport entre source, trace et carte.

Le livre s'achève sur deux cas empiriques qui exemplifient l'usage des traces numériques dans les domaines les études territoriales. Dans l'analyse de Laurent Beauguitte et Marta Severo, les traces numériques deviennent sources d'information géographique. Les auteurs proposent d'utiliser des traces médiatiques, en particulier les fils RSS des journaux quotidiens pour étudier les rapports entre individus et échelles, à travers l'exemple du *New York Times*. Cette recherche, réalisée dans le cadre du projet ANR Corpus Géomédia, met en évidence les potentialités de ces nouvelles données pour l'analyse des territoires mais elle cherche également à identifier les problèmes méthodologiques que leur usage soulève.

Enfin, Nicolas Douay et Maryvonne Prévot analysent l'impact des traces numériques sur les politiques urbaines. Comme le soulignent les auteurs, aujourd'hui les outils numériques ont contribué à la création des nouvelles dynamiques participatives de mobilisation sociale et de co-construction de la décision dans le contexte de la ville. Le cas de la plate-forme *Carticipe* à Strasbourg et à Marseille est analysé dans ce chapitre à travers des entretiens avec différents acteurs qui ont contribué à son développement et usage. Si la nature positive de ces expérimentations émerge clairement, la complexité de la cohabitation de ces innovations numériques avec d'autres dispositifs plus classiques de la participation établis par les acteurs institutionnels est également évidente.

En dépit de la différence d'approches et de perspectives, nous pouvons retracer au moins trois traits communs à ces interventions. Premièrement, une convergence vers la question des traces numériques, des données ou des *big data* – une distinction terminologique qui reste souvent gérée de manière implicite. Il nous semble que les traces représentent aujourd'hui le point de rencontre et discussion pour maintes disciplines qui s'intéressent au numérique, selon différentes perspectives – ontologiques, épistémologiques et méthodiques, anthropologiques, éthiques et politiques, pratiques, etc. Deuxièmement, tous les auteurs paraissent partager une approche ambivalente – « pharmacologique », nous pourrions dire – envers ces traces et leurs usages. La conscience des avantages de ces nouvelles sources d'information sur la réalité sociale et le territoire est contrebalancée par celle de leurs limites. La méfiance par rapport aux mauvaises pratiques liées à leur traitement – surveillance, quantification du soi, etc. – est modérée par une confiance dans les bons usages. Troisièmement, les auteurs ne s'arrêtent pas à cette constatation, mais se soucient souvent de donner une mesure précise des potentiels et des limites, des dangers et des avantages des traces numériques. En conclusion, nous pouvons dire que cet ouvrage, loin encore de constituer le point final de la recherche sur les traces numériques et le territoire, indique une direction pertinente pour ceux qui veulent poursuivre dans ce domaine.

Partie 1

La trace, les méthodes et les données

Au-delà de la critique *big data*

La recherche sociale et politique à l'ère numérique

Richard Rogers

Introduction : *big data* et le tournant computationnel

Parmi les débats actuels se pose la question de l'harmonisation de la recherche en sciences politiques et sociales avec l'ère du numérique. Comment faire face aux défis que posent Internet et le numérique, notamment les données désormais accessibles en ligne, à la recherche ? Alors que le terme *big data* connaît un succès grandissant, il semble que la recherche, tout comme l'ensemble de ce que l'on pourrait appeler le calendrier programmatique (numéros spéciaux des revues, appels à financement, titres des conférences, séminaires, etc.), soit guidée par des méthodes et des outils bien spécifiques. Pour certains, ce phénomène a pris le nom de tournant computationnel, qui désigne l'intégration des techniques informatiques et des *big data* en plein essor, aux pratiques appliquées à la recherche en sciences sociales [Berry, 2011]. Dans un premier temps, je propose d'examiner le tournant computationnel, les *big data* et les multiples critiques qui leur ont été adressées, pour ensuite les étudier indépendamment de ce que l'on pourrait appeler le « tournant numérique ». Ce dernier offre en effet certaines opportunités de recherche s'appuyant sur l'analyse mais selon une approche différente de ce que l'on a pu observer dans le tournant computationnel et les *big data*.

Le débat consiste en partie à estimer l'étroitesse des liens qui unissent le tournant computationnel à l'influence croissante des *big data*, une notion qui, dans le « hype cycle » du cabinet Gartner, a évolué du stade des « attentes exagérées » en juillet 2013 à celui du « gouffre des désillusions » un an plus tard [Gartner, 2014]. Cette descente coïncidait avec le problème de surestimation dont fut victime le site Google Flu Trends, un projet novateur de *big data* en recherche médicale et sociale grâce auquel les recherches sur les symptômes de grippes ou associés effectuées sur le moteur de recherche sont géolocalisées et utilisées pour mesurer l'activité réelle de la grippe et sa localisation. Ce service avait surévalué l'importance de la grippe (par un coefficient de deux) par rapport aux chiffres publiés par les systèmes habituels de surveillance de l'Institut national de la Santé américain et ses équivalents ; ceci avait amené les chercheurs à s'interroger sur les méthodes de mesure pratiquées sur le Web et si celles-ci ne se limitaient pas à identifier des effets médiatiques plutôt que des réalités sur le terrain [Lazer *et al.*, 2014]. A l'annonce d'une nouvelle vague de grippe, constate-t-on plus d'activité sur le moteur de recherche, sachant que la personne qui effectue les recherches est

susceptible de ne ressentir aucun symptôme ? Parmi les principales critiques adressées aux *big data* est le fait que l'on peut très bien voir un modèle là où il n'en existe pas, ce que boyd et Crawford, dans leur contribution au débat, désignent par apophénie [boyd et Crawford, 2012].

Big data : points de vue critiques

L'avènement des *big data* est censé avoir entraîné une rupture, notamment à cause du remplacement de l'interprétation par la recherche de modèles [Savage et Burrows, 2007 ; Watts, 2007 ; Lazer *et al.*, 2009]. Cela signifie que dans une démarche type de recherche, le prélèvement d'un échantillon (en sciences sociales) ou l'identification d'une collection (en sciences humaines) serait remplacé par une collecte de données par quadrillage systématique, dont la finalité n'est pas simplement la constitution exhaustive d'un mais plutôt de plusieurs ensembles de données, de grande taille et collectés selon plusieurs méthodes, notamment la capture en temps réel. L'apprentissage automatique et l'analyse des réseaux formeraient les points de départ de telles démarches analytiques. On pratiquerait en effet l'analyse non pas par une lecture « de près » mais par une lecture « à distance », ou topologique, qui élèverait, en points de mire des résultats, la cooccurrence de mots, les tendances du moment, les prescripteurs, ainsi que l'anomalie et le hors-norme [Piper, 2013]. Franco Moretti, qui a théorisé la lecture à distance, a montré que la lecture de près produit un canon qui a pour conséquence de limiter la portée de ce qui est lu, et encourage, à l'inverse, à apprendre à ne pas lire [Moretti, 2005]. En utilisant ces techniques, il serait possible de soumettre à l'analyse une plus grande production, et non plus seulement des œuvres canoniques.

Dans le domaine des *big data*, l'infrastructure computationnelle et les moyens de la préserver deviennent des obstacles à la saisie scientifique. On pourrait, d'une certaine manière, comparer l'investissement à fournir à celui d'un grand moteur de recherche ou d'une chaîne d'information. Il faut pouvoir faire preuve d'équipes dédiées de techniciens-chercheurs et de programmeurs scientifiques capables d'entreprendre une telle tâche, ce qui implique une expertise dans le traitement de données et une certaine habileté dans l'utilisation de logiciels d'analyse et de leurs nombreux paramètres. En d'autres mots, il n'est plus possible d'exploiter des ensembles de données au moyen d'un tableur ou « à la main ».

Ceci a des conséquences directes sur les systèmes réputationnels appliqués au domaine scientifique. En basculant vers des infrastructures d'envergure avec leurs équipes d'analystes et de techniciens, l'écart entre supra-science et petite-science se creuse à nouveau, au sein même, cette fois-ci, des sciences humaines et des sciences sociales [Price, 1963]. Avec l'avènement du tournant computationnel et de sa vision orientée vers les supra-sciences, la valeur d'une étude se mesure à travers le prisme

des dispositifs de capture de données employés, ainsi que par la taille et la complexité de ces données. L'application d'une méthode et d'un procédé, la combinaison de la force brute de l'équipement à la finesse des paramètres choisis, sont ici capitales.

Les résultats des analyses produisent par ailleurs des modes de connaissance à distance, à savoir des compétences particulières dans la compréhension de clusters graphiques et autres complexités visuelles. On peut comparer ces nouvelles connaissances chez les scientifiques à celles que possèdent des gestionnaires de systèmes d'exploitation, comme les contrôleurs aériens ou les analystes d'images satellites [Rochlin, 1997 ; Parks, 2005].

En collectant en continu des données en temps réel, faire de la recherche reviendrait à observer une sorte de veille médiatique à l'aide d'un dispositif de capture en « back-end » mais aussi d'un tableau de bord en « front-end ». L'interface de ce tableau de bord, à partir duquel on peut gérer à la fois la réception des données et les résultats d'analyses, offre de nombreux modes de visualisations, par exemple les graphiques, cartes et arbres de Moretti, mais également des cartes géographiques qui pourraient nous renseigner sur les points de densité de l'attention. La courbe d'apprentissage pourrait alors inclure des informations comme les dimensions des tuiles et les noms des fournisseurs de tuiles qui produisent les carrés utilisés pour la composition de la carte.

Lorsque l'on s'intéresse aux *big data*, certaines questions connexes peuvent se révéler très importantes. En tout premier lieu, on ne peut ignorer le vaste débat sur la pertinence ou non d'analyser les données personnelles à l'aide des *big data*, sachant qu'on a peu de chance d'obtenir une autorisation d'utilisation des données et qu'il est tout à fait délicat de se substituer à cette autorisation en s'appuyant sur les conditions d'utilisation de l'entreprise. Si un cas d'éthique est jugé acceptable, il faut alors mettre en place des stratégies qui, tout en n'exigeant pas d'autorisation, ne soient pas préjudiciables. La solution de l'anonymat ne convient pas forcément ; cela tient en partie à la situation à laquelle furent confrontés les chercheurs en données d'AOL, lorsque des journalistes et des défenseurs du droit à la vie privée ont pu identifier certains utilisateurs du moteur de recherche après que leurs recherches des six derniers mois ont été rendues publiques [Sweeney, 2000 ; Hansell, 2006]. On peut également citer la technique du *cloaking* ou de dissimulation, plus communément appelée pseudonymisation, par laquelle des identifiants artificiels sont insérés à la place des véritables noms. Une autre manière de répondre aux questions éthiques qui émergent de l'analyse des *big data* personnelles est de les agréger et de les dépersonnaliser : c'est le cas, par exemple, lorsque l'on recueille les listes des intérêts postés sur leur page Facebook par les amis des candidats à la présidence américaine (ou de n'importe quelle autre personnalité publique qui possède une page officielle) dans le but de les observer et d'en déduire certains rapports de cause à effet, comme par exemple l'influence des guerres entre cultures ou les préférences médiatiques des amis selon leur orientation politique à gauche ou à droite.

Aujourd'hui, les critiques des *big data* pointent une tendance à davantage se fier à des flux et des signaux de données disponibles (par exemple les APIs des réseaux sociaux) plutôt qu'à appliquer la méthode usuelle qui consiste à en chercher de nouveaux mais qui demande davantage d'efforts [Vis, 2013]. Dans les fondamentaux de la recherche, les données auraient-elles remplacé une vision qui privilégiait les problèmes à résoudre ? Par exemple, quelqu'un qui voudrait géolocaliser des poches de haine ou d'intolérance doit-il s'engager sur la piste potentiellement prometteuse des tweets géotaggés, ou doit-il prendre en compte des données nouvelles et les moyens de les recueillir ? Dans ce cas, c'est la disponibilité immédiate des données qui déclenche le questionnement scientifique et la recherche qui en découle, et non une question sociétale en soi.

Souvent, lorsque les *big data* sont impliquées, on se tourne vers les grandes compagnies de médias et de communications pour récupérer des groupes de données, données qui réorientent ensuite vers d'autres intérêts bien précis (par exemple, l'utilisation des données mises à disposition par les compagnies de téléphone mobile pour retracer les différents types de trajets empruntés quotidiennement). Une autre manière de s'assurer un accès facile à des données exploitables est de s'asseoir, littéralement, à l'un des postes de travail d'un laboratoire d'une grande compagnie en tant qu'invité ou collaborateur chercheur. On peut alors mener la recherche sur place mais il est impossible de quitter le bâtiment avec les données, ou de les sauvegarder sur un serveur à distance.

Ainsi, les modalités d'accès aux *big data* deviennent elles-mêmes sujets d'études, y compris pour celles qui bénéficient d'un accès public, comme les archives Web des bibliothèques nationales et bientôt l'archive Twitter du Congrès américain. Sur un mode comparable au laboratoire d'une grande compagnie, le chercheur doit souvent se trouver au sein même de la bibliothèque pour pouvoir accéder aux données, qu'il ne peut emprunter comme il le ferait avec des livres. La différence entre nouveaux médias et médias traditionnels est ici flagrante et l'enjeu des débats sur les modes de production de la connaissance à partir des *big data* et des techniques de lecture à distance, apparaît ici pleinement [Schulz, 2011 ; Allison *et al.*, 2011]. Non seulement faut-il être, si l'on peut dire, proche des données pour pouvoir les lire à distance ; il faut encore obtenir le droit de les approcher et d'y accéder physiquement.

Il faut encore ajouter à cela le temps passé à travailler sur les données, une mission très longue, surtout lorsque l'on doit rester sur place, avec les données. Comme il a été mentionné plus haut lorsqu'ont été évoqués les phénomènes de pseudonymisation et de dépersonnalisation des données, avant même l'analyse des données, il faut parfois entreprendre des démarches nouvelles en terme de prévention, des démarches récemment apparues qui répondent aux actes de désanonymisation initiés par des tiers (journalistes et défenseurs de la vie privée notamment). Il est important de se préparer à la montée de ce que l'on appelle la science de la ré-identification,

ou l'ensemble des techniques qui permettent de rétablir l'identité de personnes délibérément brouillée [Ohm, 2010]. On a, par conséquent, ajouté au nettoyage des données une étape préparatoire supplémentaire.

Pour résumer, la critique des *big data* a des points d'ancrage épistémologiques, esthétiques et éthiques, à commencer par une rupture conceptuelle en faveur de la reconnaissance de modèles, un mode analytique de l'apprentissage automatique, des préférences en termes de graphiques, cartes et arbres, et tout ce qui touche à la manière de gérer les données de sujets humains. Les *big data* divisent aussi, en créant des structures réputationnelles par l'intervention, à chaque étape de la recherche, d'une perspicacité – ou cécité – computationnelle : collecter, accéder, traiter, analyser, produire et présenter les données.

Les *big data* peuvent être commercialisées. De nouveaux marchés de données se développent, ce qui implique qu'il faille parfois payer pour mener des recherches sur les données [Puschmann et Burgess, 2013]. Pour accéder aux données historiques de Twitter, fournies par Gnip (lui-même détenu par Twitter), on peut souscrire à un abonnement ou déposer des demandes ponctuelles sur des collections de tweets, ce qui, dans les deux cas, sépare les bases de données riches des bases de données pauvres. Sifter, l'interface de Texifter qui propose l'accès à des données historiques de Twitter, offre une estimation du coût de la transaction ; par exemple, fin 2014, une demande portant sur l'accès à toutes les données historiques des tweets contenant un hashtag encore actif depuis 2007 revenait à 20 000 USD.

Du tournant computationnel au tournant numerique

Une autre manière, plus subtile, de caractériser l'invasion des ordinateurs, qu'incarnent notamment ces ordinateurs portables bardés d'autocollants et les Hacking Workshop, pourrait être le tournant numérique, où l'étude et les méthodes de la culture numérique influencent la recherche dès lors qu'elle utilise les données (en ligne), les logiciels d'analyse et la visualisation infographique. Faire la distinction entre le tournant computationnel (des *big data*) et le tournant numérique permet de ne pas céder à une compréhension monolithique, ou unitaire, de l'évolution des besoins en recherche politique et sociale à l'ère du numérique [Lovink, 2014].

Au sein même du tournant numérique, il existe tout un panel d'approches qui englobent les humanités numériques, les sciences sociales numériques et celles des médias numériques (qui forment toutes les « études numériques » ou *digital studies*) et dont les engagements et positionnements ontologiques et épistémologiques se distinguent à première vue nettement les uns des autres. Je vais tâcher ici de

contextualiser et d'examiner certaines pratiques de la recherche numérique observées en humanités numériques (analytique culturelle et culturomique), en sciences sociales (cybermétrie et mesures alternatives d'impact ou *altmetrics*), et dans des études croisées (méthodes numériques), et de montrer, à travers quelques exemples, en quoi elles peuvent contribuer à la recherche en sciences politiques et sociales [Manovich, 2011 ; Michel *et al.*, 2010 ; Priem *et al.* 2010 ; Rogers, 2013]. Je m'intéresserai particulièrement à ce que l'on a appelé les méthodes et techniques quali et quanti du tournant numérique, ainsi qu'aux méthodes de recherche utilisées lors d'un « data sprint », en les envisageant séparément du phénomène *big data*.

En premier lieu, on peut différencier l'ensemble des méthodes utilisées en sciences numériques les unes des autres, selon qu'elle utilise un type de document plutôt qu'un autre, ou une méthode bien précise ; ceci revient à distinguer les études qui utilisent des documents et des méthodes issus de la numérisation, les natifs numériques, ou une combinaison des deux[1]. Sur la question des documents, les chercheurs en humanités numériques, tout particulièrement, se voient de plus en plus offrir de nouveaux corpus de recherche alors que les bibliothèques nationales, les archives et les musées livrent régulièrement des fournées fraîches de matière numérisée et indexée. On parle même de crise dans certains domaines de recherche en humanités numériques, car l'accueil réservé par les chercheurs à ces nouveaux documents n'a pas été à la hauteur de cette ruée vers la numérisation. Par exemple, les archives Web sont encore rarement utilisées à des fins scientifiques, si l'on en juge par le nombre de publications citant des collections archivées de données Web [Dougherty *et al.*, 2010 ; Thomas *et al.*, 2010]. De plus, certaines interrogations subsistent quant à l'équipement et la formation des enseignants-chercheurs et des étudiants, là où l'enjeu essentiel porte à la fois sur la pertinence et la manière d'enseigner la programmation (ou le codage) [Manovich, 2015].

Pour les chercheurs en sciences sociales numériques, les données numériques natives (issues du Web) pourraient présenter un plus grand intérêt que les objets patrimoniaux numérisés, dans la mesure où les données sont utilisées pour étudier les tendances, les opinions, les rumeurs, les prises de position, etc., ainsi que l'avait signalé en 2009 un groupe de chercheurs en sciences sociales qui s'était penché sur le tournant computationnel dans sa discipline [Lazer *et al.*, 2009]. Reste à savoir si vraiment, comme certains se le sont demandé, on peut évaluer l'opinion publique en étudiant les tendances à partir, par exemple, de données recueillies sur Twitter. Existe-t-il une corrélation entre l'humeur détectée sur Twitter et le comportement des « esprits animaux » sur le marché boursier [Bollena *et al.*, 2011] ? Peut-on inférer l'incidence de maladies et de troubles médicaux, des tweets touchant à la santé [Mitchell et Hitlin, 2013] ? Le problème se pose dans les mêmes termes que pour le Google Flu Trends : étudie-t-on ici une tendance sociétale ou le fonctionnement des réseaux sociaux ? Dans quels cas l'étude de données Twitter ne concerne-t-elle

1 Voir également Rogers [2009].

que Twitter, et dans quels cas (ou jusqu'où) est-il possible d'étudier grâce à Twitter des phénomènes sociétaux « à l'état naturel » ? A l'occasion d'un concours organisé en 2013 par le Centre américain pour le contrôle et la prévention des maladies (*US Centers for Disease Control*, CDC), une série d'études avait été publiée qui avait montré comment l'évolution de la grippe pouvait être suivie sur Internet non seulement grâce au moteur de recherche de Google mais également par le type de pages vues sur Wikipédia et de tweets géotaggés sur Twitter. On pourrait qualifier cette méthode qui consiste à obtenir des résultats à partir de données disponibles en ligne d'« inférence en ligne » : les résultats sont souvent accessibles sur des interfaces type tableaux de bord, avec des cartes montrant les zones d'activité, comme l'a fait par exemple le lauréat du concours du CDC, « Prédire la prochaine saison de grippe, un site de prévision des maladies infectieuses (grippe) de l'Université de Columbia » [CPID, 2015]. Si l'algorithme de prévision s'appuie sur la combinaison de données de surveillance compilées par le CDC avec l'aide d'un Google Flu Trends révisé (et réhabilité), l'authenticité des résultats n'en repose pas moins sur les régimes de surveillance traditionnels tels que les admissions à l'hôpital, ce qui leur donne davantage de crédibilité.

En termes de méthode, on pourrait considérer que les sciences sociales et humanités numériques numérisent la méthode, dans le sens où elles déplacent leurs instruments sur le Web (enquêtes, sondages, etc.) ou les intègrent à des logiciels (en mesurant et éditant les propriétés formelles des objets d'art, comme dans le cas du logiciel ImagePlot). On pourrait pousser le raisonnement un peu plus loin et envisager la numérisation des concepts ou des projets de recherche, adaptant ainsi au Web les méthodes de recherche pratiquées en sciences politiques et sociales (par exemple en étudiant l'expression de la sphère publique dans les forums, en s'attaquant aux espaces réservés aux commentaires pour y trouver des traces de débat public, ou en étudiant Twitter comme source de l'opinion publique).

Faire la distinction entre des documents numérisés et des documents numériques natifs, entre méthode et concept, revient à imposer des distinctions ontologiques entre ceux qui « appartiennent au médium » et ceux qui l'ont intégré à la suite d'une migration [Blood, 2002]. Les blogs, que l'on considère comme appartenant au Web, sont dans ce cas des numériques natifs, tandis qu'un livre numérisé, disponible sur Google Books, est un nouvel arrivant numérique, soit un document numérisé. Les pages Web qui ne peuvent être imprimées mais seulement capturées à l'écran sont une autre manière de conceptualiser la distinction entre objets du médium et ceux qui en dérivent [Latour, 2004].

La distinction entre natifs numériques et documents numérisés peut également s'appliquer à la méthode employée dans les sciences humaines et sociales. On trouve, d'un côté, des méthodes qui ont effectué une migration vers le Web – les enquêtes en ligne par exemple –, et de l'autre, des méthodes spécialement créées pour une utilisation en ligne, comme le PageRank de Google (un classement

qui privilégie un site Web plutôt qu'un autre) ou l'EdgeRank mis en place par Facebook (qui privilégie un ami plutôt qu'un autre selon votre affinité avec lui).

On peut donc placer les différentes approches en recherche numérique sur une matrice binaire (ou, si l'on préfère, les disposer sur un graphique en éventail, ou bien les faire défiler horizontalement à l'aide d'un curseur), où l'on verrait apparaître quels types de documents sont davantage sollicités pour les données (numérisés ou natifs) et quelles méthodes sont imitées ou natives (voir Tableau 1).

Enfin, en lieu de la numérisation des concepts comme cela a été évoqué plus haut, existe un discours spécifique à l'usage du Web, que l'on entend aussi bien dans la presse professionnelle que dans l'utilisation des logiciels critiques, un langage dont on pourrait tirer une utilisation analytique : le *spam*, les *creative commons*, la wikification, le *versioning* – ou gestion de version –, le *malware* – ou logiciel malveillant –, le codage, la plate-forme, le profil, etc. Des études telles que celles de Lawrence Lessig et Eben Moglen ont expressément souligné l'utilisation, dans le discours juridique des nouveaux médias, des mots formés à partir du préfixe « re- », des termes tels que le *remix*, le *reuse* – ou ré-utilisation, la redistribution, la révision et le recyclage – *redistribute*, *revise* et *recycle* –, qui apparaissent comme autant de défis lancés par les médias numériques aux lois et mœurs existantes. De façon similaire, l'étude des natifs numériques, ou ses enjeux, qui est mise en avant dans la notion même de méthodes numériques, propose de réorienter la méthodologie du médium et des concepts qui lui sont associés, et de les appliquer à la recherche en sciences sociales et politiques. Cette variante du tournant numérique invite en effet à étudier non seulement ce qui est numérisé mais également le numérique natif, en se servant des données et en suivant un raisonnement qui reprenne les méthodes du médium. Comment faire pour réorienter, dans le cadre de la recherche en sciences politiques et sociales, les techniques et notions numériques vers une utilisation en ligne ?

Pour répondre à cette question, je vais m'intéresser à un nouvel espace d'expression politique (Facebook) et développer une approche analytique à la fois qualitative et quantitative (ou quali-quanti) qui utilisera les données recueillies sur Facebook (mentions J'aime, partages, commentaires, mentions J'aime sur les commentaires) afin d'étudier l'engagement politique, une préoccupation fréquente en sciences politiques mais également un terme utilisé par les analystes spécialisés en médias. Cette méthode numérique se fonde sur un mode opératoire appelé analyse des contenus en réseau (*networked content analysis*) qui associe le comptage à l'interprétation du contenu le plus sollicité au sein d'un réseau d'utilisateurs.

Disons-le d'emblée, les méthodes numériques fonctionnent, que ce soit pour l'analyse des contenus en réseau ou d'autres protocoles de recherche, et on voit de plus en plus se développer un nouveau format de recherche appelé « *data sprint* ». Les *data sprints*, qui partagent quelques similarités avec les hackathons, les *barcamps*, et autres formes expérimentales de workshops reposant sur le partage

de compétences, opèrent sur un format court, en concentrant l'exécution pratique du travail de recherche sur une durée courte, par exemple une semaine. Des équipes composées d'experts en la matière, analystes, programmeurs et designers d'information, travaillent ensemble à la production d'un projet d'études sous la forme d'un court rapport et d'une série de graphiques. Il arrive souvent que plusieurs projets soient conduits en parallèle et chacun des projets est présenté à tous les autres groupes à la fin de cette semaine de « sprint ». C'est une forme de « science rapide » ou *fast science*, qui n'est ni de la supra-science ni de la petite-science, mais qui fait partie des pratiques apparues avec le tournant numérique.

Données numérisees et données numériques natives

Lorsque l'on entreprend de créer un corpus de documents numérisés, on peut appliquer les mêmes principes que ceux utilisés pour la constitution de « bonnes » données ou données fiables (« *good data* »), à savoir la numérisation de données « propres » et entières, couvrant toute la durée de vie de leur production [Borgman, 2009]. Dans le domaine des humanités numériques, la collection complète des œuvres (connues) d'un artiste ou celle de toutes les couvertures d'un magazine prescripteur tels que le *Time Magazine* représentent des données fiables et font l'objet d'études en analytique culturelle, comme je le montrerai plus loin. Dans le domaine des sciences sociales, les mesures médiatiques telles que les taux d'audience mesurés par Nielsen et les tirages des journaux estimés par le BPA (*Business Publications Audit*) ont recours à d'autres principes de fiabilité des données (comme l'échantillonnage d'une population donnée).

Toutefois, on observe que le Web ne dispose pas d'une telle instance de comptabilisation [Graves et Kelly, 2010]. Prenons Google et le nombre de résultats estimés lors d'une recherche sur le Web : comme l'a révélé un cas bien connu, la requête associée à un terme générique comme [voitures] (*cars*) produit moins de résultats que celle, plus ciblée, de [voitures d'occasion] (*cars – used*) [Sullivan, 2010]. En effet, les données numériques natives ne partagent pas nécessairement les mêmes propriétés que les « *good data* » car le Web de manière générale, et les plates-formes de réseaux sociaux en particulier, sont des médias instables, c'est-à-dire éphémères, et en perpétuel mouvement. Constamment de nouvelles caractéristiques (et par conséquent des types de données) sont ajoutées ou rendues obsolètes.

La question de la fiabilité des méthodes de mesure du Web dans une étude longitudinale pose donc problème. Le bouton « J'aime » de Facebook en est un très bon exemple : sa date de création remonte à 2010 et, par conséquent, toute mesure antérieure de l'activité de Facebook ne pourra prendre en compte cette caractéristique, qui est l'une des plus populaires du site. De plus, de sa création à la fin de l'année 2011, l'utilisation du bouton a augmenté de manière progressive si bien que les analyses sur cette période risquent d'entraîner une sous-évaluation

des données (« *underreporting* »), même en cas de pic d'activité [Richmond, 2011]. De la même manière, les résultats de recherche sur Google font preuve d'une certaine volatilité dans le temps. Après d'importantes mises à jour de certains de ses algorithmes par Google (Panda en 2011 par exemple), des sites internet ont vu leur classement baisser, notamment ceux qui présentaient liens « spammy », fermes de spam ou répétiteurs de contenus (« *content repeaters* ») en nombre trop important [Singhal et Cutts, 2011]. Il y a fort à parier que si l'on étudiait l'évolution dans le temps des classements des résultats des moteurs de recherche, on verrait à quel point des algorithmes instables produisent des positionnements fluctuants.

DATA	METHOD: DIGITISED	METHOD: NATIVELY DIGITAL
DIGITISED	▸ Culturomics ▸ Cultural Analytics	▸ Altmetrics
NATIVELY DIGITAL	▸ Webometrics	▸ Digital Methods

Tableau 1 : Cinq manières d'étudier les humanités numériques et les sciences sociales numériques représentées selon les types de données et méthodes privilégiés

Les études numériques

Ces dix dernières années, les méthodes et techniques déployées dans le cadre de la recherche numérique (qui utilise à la fois des données numérisées et des données en ligne), ont été énoncées selon une large gamme de descripteurs, notamment : l'analytique, les indicateurs, les sciences -nomiques ou toutes méthodes connexes, projetant ainsi tout un panel de nuances sur l'objet mesuré. En recherche numérique, on associe naturellement l'utilisation de l'analytique à l'industrie de surveillance des nouveaux médias, qui est de plus en plus impliquée dans la surveillance des réseaux sociaux et l'écoute sociale, comptabilisant et transformant en indicateurs l'activité des utilisateurs. On capture et on analyse les mentions, puis on en déduit une signification afin d'obtenir « une vision opérationnelle », pour reprendre les termes d'un fournisseur de logiciels [Adobe, 2014]. Dans la plupart des cas, la recherche de la signification concerne l'impact et la mesure du sentiment. Un nom, un mot-clé, une marque ou un autre terme s'est-il/elle propagé(e) ? S'est-il/elle diffusé(e) au-delà d'un public captif (par exemple les followers d'un réseau social), trouvant ainsi

un écho en dehors de cette « bulle » ? Cette diffusion a-t-elle été bienveillante ? On peut également mettre en place une veille sur la compétition entre des noms, des mots-clés, des marques et autres termes, et voir dans quelle mesure une campagne (ou un programme) trouve davantage d'écho (parmi les prescripteurs) que sa contre-campagne (ou contre-programme) [Latour, 1991].

Les métriques (*metrics*) sont des normes de mesure ; leur nomenclature provient des techniques de comptage utilisées notamment en bibliothèque et en sciences de l'information, telles que la bibliométrie et la scientométrie. Ils s'intéressent au facteur d'impact, à la pertinence et à l'influence, en d'autres termes non pas uniquement à la force brute mais à la force relative et à son endurance. Sur les réseaux sociaux et sur tout autre environnement numérique personnel quantifié (par exemple les logiciels d'évaluation d'un programme fitness), les métriques comptabilisent et communiquent l'intensité de l'activité – les « J'aime » sur Facebook par exemple –, et ont été notamment critiquées pour leur incitation à « "vouloir toujours plus", poussant les gens à vouloir toujours plus de "J'aime", plus de commentaires, et plus d'amis » [Grosser, 2014].

L'utilisation du suffixe -nomique est moins directement liée à la question des relations entre l'industrie du Web et les sciences ; il renvoie à la « loi », dans le sens des lois de la nature, et fait donc allusion aux fondamentaux de la vie et la recherche fondamentale. Il partage avec le terme « méthodes » une approche épistémologique illimitée. Quelle que soit notre progression dans le travail, et quelle que soit la démarche adoptée, les méthodes soulignent la construction par étapes d'une procédure ou d'un protocole de recherche. Si l'on applique cette description aux méthodes numériques, celles-ci pourraient embrasser l'ensemble des procédures utilisées pour l'étude des documents numériques, même si le terme désigne plus spécifiquement les méthodes utilisées en ligne pour l'étude des données du Web. Je reviendrai à cette question après un bref examen de l'analytique culturelle, de la culturomique, de la cybermétrie et des métriques ; après les avoir présentées, j'offrirai quelques pistes pour les adapter à la recherche en sciences politiques et sociales.

La première de ces approches en humanités numériques, l'*analytique culturelle*, a souvent comme objet d'étude des collections numérisées (par exemple l'ensemble des couvertures historiques de *Time Magazine*, toutes les peintures de Rothko ou de Mondrian, etc.), et s'intéresse à leurs propriétés formelles et la manière dont elles changent avec le temps. Son logiciel attitré est ImagePlot, qui regroupe des images selon leurs propriétés formelles telles que la luminosité et la saturation (voir Figure 1). On appelle le produit de cette analyse un « espace de style » : « les images visuellement similaires sont dites proches ; celles qui sont différentes se retrouvent éloignées les unes des autres » [Manovich, 2011]. Ce terme s'inspire de la notion de styles utilisée en histoire de l'art, qui permet de regrouper et de réunir en catégories des œuvres d'art.

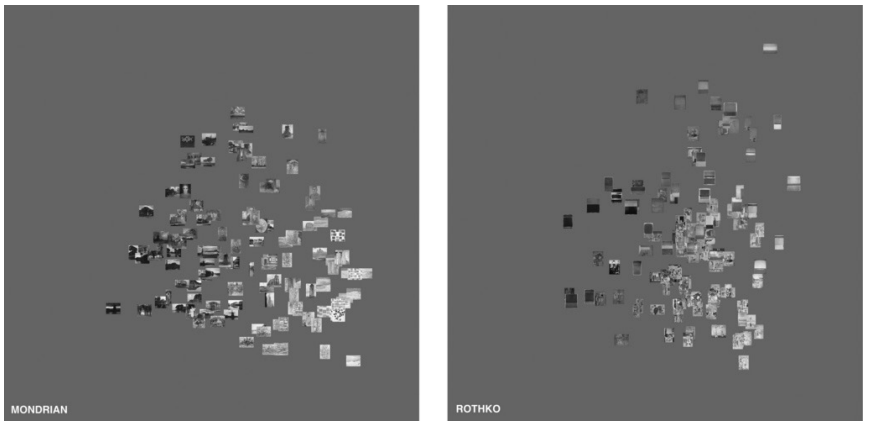


Figure 1 : Rothko contre Mondrian. Produit par le logiciel ImagePlot et Cultural Analytics. Comparaison des espaces de style. L'axe des abscisses montre la luminosité, l'axe des ordonnées la saturation. Source : Software Studies Initiative, 2011

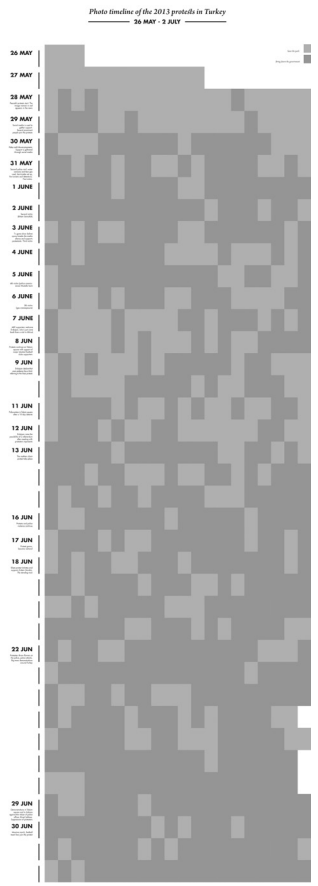


Figure 2 : Caractérisation d'image sur les premières images apparues sur Google Images lors d'une recherche sur le terme [Gezi], selon « save the trees » (« sauver les arbres », en vert) ou « bring down the government » (« faire chuter le gouvernement », en rouge), juin 2013. © Digital Methods Initiative, Amsterdam, 2013

Cette notion d'espace de style a été utilisée dans l'analyse des *selfies*, ces autoportraits photographiés à l'aide d'un smartphone et postés ensuite sur Instagram par des utilisateurs dans plusieurs villes. Le projet avait pour but d'analyser les propriétés formelles des portraits, afin de les regrouper et d'en extraire certaines constantes comme la pose (inclinaison de la tête), le sexe, l'âge et l'expression du visage, dans les villes de Berlin, Moscou, Sao Paolo, Bangkok et New York ; parmi les conclusions, on peut mentionner le fait qu'il y ait nettement plus de femmes que d'hommes qui prennent des selfies, qu'à Sao Paulo on pose avec beaucoup de sérieux, avec un angle d'inclinaison de la tête à environ 17 degrés. Certaines villes (Sao Paulo et New York) affichent plus de bonne humeur que d'autres (Berlin et Moscou, où les auteurs de selfies sont plus âgés). L'étude ne va pas jusqu'à mesurer la teinte et la saturation pour en inférer des conclusions sur la culture de l'exposition en ligne chez les jeunes.

Comment appliquer l'analytique culturelle à la recherche en sciences politiques ? On peut citer en exemple une étude qui portait sur des clichés pris lors de manifestations et recueillis sur des médias en ligne, et pour lesquels on pouvait reconstituer une chronologie des événements ; cela avait été notamment le cas avec des images prises au parc Gezi à Istanbul durant les manifestations de mai-juin 2013. En utilisant la technologie ImagePlot, on assiste à la transformation qui a traversé cette « révolution des arbres » en Turquie ; comme le dit un témoin, « la transformation de l'espace public en espace privé montre pourquoi le but de l'occupation du parc Gezi n'était pas seulement de sauver des arbres mais de sauver la démocratie turque » (voir Figure 2) [Turkey EJOLT team, 2013]. La couleur verte dominante décline progressivement et cède la place aux images de manifestants aspergés de gaz lacrymogènes puis, plus généralement, à la lutte pour les droits.

La *culturomique*, également pratiquée par les humanités numériques, est une discipline qui interroge (à partir de Google Ngram Viewer) la base de livres numérisés de Google sur des mots précis, faisant ainsi ressortir des tendances culturelles ou sociétales ; elle se concentre essentiellement sur les ouvrages en anglais parus entre 1800 et 2000, bien que l'on trouve aussi des livres dans d'autres langues [Michel *et al.*, 2011]. On obtient, sous forme de graphiques, la variation de fréquence d'utilisation de mots-clés à travers la période mentionnée ; ces graphiques rappellent (à la fois méthodologiquement et visuellement) les premières versions du dispositif Google Insights qui montrait les statistiques associées à des recherches d'utilisateurs par mots-clés sur les moteurs de recherche. Ces recherches peuvent revêtir une nature politique, et certaines peuvent diriger les utilisateurs vers des sites se situant plutôt à gauche ou plutôt à droite. Ainsi, dans la course aux élections présidentielles américaines en 2012, ceux qui rentraient le terme « *obamacare* » étaient orientés en majorité vers des sites identifiés à droite, alors qu'une recherche contenant les mots-clés « *obama student loan forgiveness* » (« obama emprunt étudiant dispense ») vous amenait vers des sites identifiés à gauche (voir Figure 3) [Borra et Weber, 2012]. La recherche sur Internet comme sujet de recherche – ce qui est une façon de décrire la culturomique – ou l'analyse

des termes de recherche peut également comprendre la géolocalisation des utilisateurs, et donc mener à un travail sur l'utilisation des mots selon la localisation de leurs utilisateurs. Sur le même mode que ce qui a été fait pour l'évaluation et la localisation de la grippe sur Google Flu Trends, on pourrait très bien géolocaliser les discours de haine (à partir d'une recherche de termes linguistiques précis) et ainsi observer de manière longitudinale ses constantes ou ses fluctuations.

politicalinsights
obama
Search
(Example: try to search president)
Top political queries grouped by leaning, containing obama
Next
Ordered by left leaning
1. obama accomplishments
2. obama's accomplishments
3. obama student loan forgiveness
4. obama press conference
5. list of obama accomplishments
Ordered by center leaning
1. obama polls
2. obama poll
3. obama's aunt
4. obama poll numbers
5. barack obama's religion
Ordered by right leaning
1. cost of obama's trip to india
2. obamacare
3. obama trip to india cost
4. obama affair rumors
5. obama collective salvation
Political Insights by Yahoo! Labs displays the proportion of times a search query results in a click-through on a political blog of a particular leaning. Blue stands for left-leaning blogs, grey for center, and red for right-leaning blogs. For more information see the about page.
Copyright © 2011 Yahoo! Inc. All rights reserved. Help - Privacy - Terms of Service

Figure 3 : Political Insights, Yahoo! Labs, montrant les recherches des sympathisants de gauche et de droite ayant pour sujet Obama, 2011. Source : Borra et Weber, 2012

En sciences sociales numériques, la *cybermétrie* consiste à appliquer des méthodes d'analyse de citations en utilisant des liens Internet (principalement) de la même manière que s'ils étaient des citations extraites d'ouvrages universitaires, et en traitant le lien Internet comme un gage de crédibilité ou de mesure d'impact [Thelwall *et al.*, 2005 ; Ackland, 2013]. On retrouve l'approche cybermétrique dans des logiciels tels que Issuecrawler ou VOSON, qui fouillent méticuleusement (ou «*crawl*») les sites Web, identifient des liens et représentent ces relations sous la forme de graphes de réseaux : ils en montrent ainsi les caractéristiques, y compris la centralité ou la périphéralité d'un ou plusieurs acteurs. Elle peut également mettre en évidence une stratégie en ligne, comme l'ont montré les graphes de réseaux publiés par Issuecrawler suite à la campagne en ligne d'Obama en 2008[2]. Le caractère exceptionnel de la forme étoilée du réseau s'explique par la stratégie de mise en relation par laquelle s'est illustrée la campagne d'Obama en 2008 (voir Figure 4). Le cœur du réseau est formé par le site http://barackobama.com et ses sous-sites, tels que http://latinos.barackobama.com, http://faith.barackobama.com et http://students.barackobama.com. La périphérie est constituée en majeure partie par des sites de réseaux sociaux consacrés à Obama, ainsi que par les pages du candidat

2 Voir également Foot and Schneider [2006] ; Krippendorff [2012].

sur LinkedIn, Facebook, Flickr, etc. Le réseau écarte par ailleurs d'autres sites et ne fait donc pas état des campagnes – Web – populaires influencées par les nouveaux médias, comme celle de Howard Dean en 2004, qui avait permis aux utilisateurs de créer leurs propres histoires lors des rencontres ; il s'efforce davantage de montrer une approche centralisée, concentrée sur le message à délivrer [Rogers, 2005].

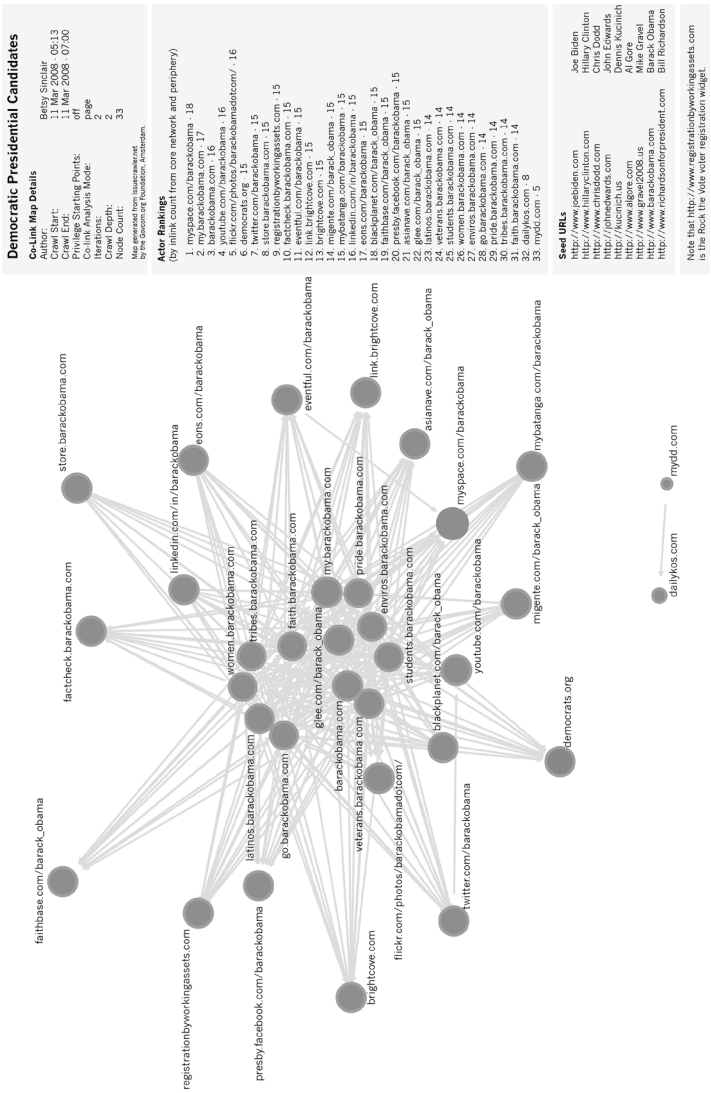


Figure 4 : Graphique produit par Issuecrawler sur les interconnections entre les différents sites Internet consacrés à Obama, 2008. Source : Issuecrawler.net, © Govcom.org Foundation, 2008, mentionné dans Krippendorff, 2012

Les mesures alternatives d'impact ou *altmetrics* diffèrent de celles issues de la scientométrie traditionnelle dans le sens où elles comptabilisent les citations des ouvrages universitaires publiés non dans des revues mais sur les blogs, Twitter,

ou tout autre réseau social universitaire ou plate-forme de citation de type http://academia.edu ou Mendeley. La comptabilisation (et l'interprétation) de références bibliographiques sur les réseaux sociaux participe d'une approche analytique plus large de la matière substantielle et des responsabilités de la source sur une question d'actualité ou un réseau idéologique, observé par exemple sur Facebook ou Twitter.

On peut prendre comme exemple les contenus les plus référencés sur Twitter (dans ce cas précis, ceux qui renvoient le plus à des pages Web) par les fonctionnaires hollandais rattachés à des ministères. Il s'avère que ces fonctionnaires tendent naturellement à suivre les actualités, les hommes ou femmes politiques, ainsi que les observateurs des nouveaux médias et de la vie politique, plutôt que de simples citoyens (voir Figure 5). De plus, le sujet le plus référencé est l'utilisation, par les fonctionnaires, des nouveaux médias et les actions et campagnes innovantes sur le Web, ce qui signifie que le contenu partagé est avant toute chose auto-référentiel et centré sur les médias plutôt qu'un sujet d'actualité.

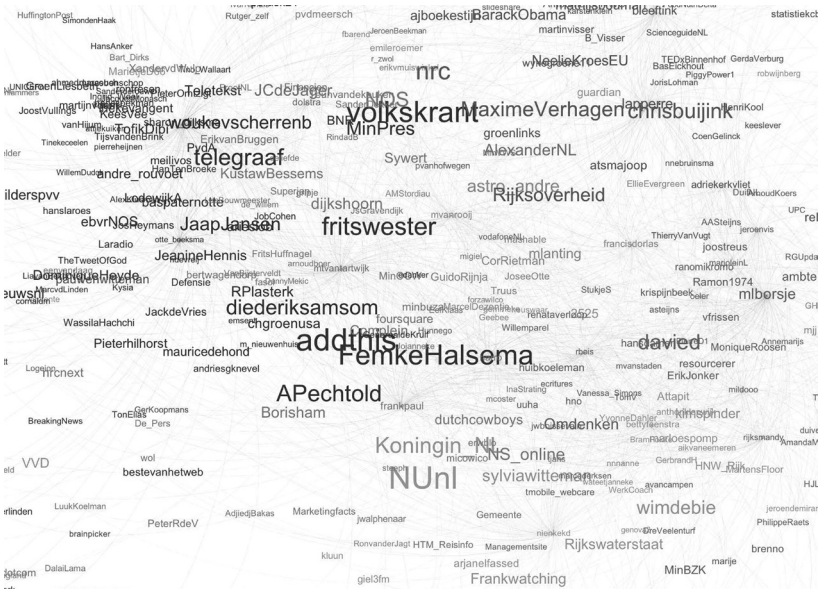


Figure 5 : Réseau élargi des follow-followers des fonctionnaires ministériels hollandais, mars 2013. Données recueillies par TCAT, DMI Amsterdam, et représentation graphique avec Gephi. Source : Baetens et al.*, 2013*

Comme nous l'avons déjà mentionné, pour certains, le terme « méthodes numériques » couvre l'ensemble des technologies du tournant numérique que nous avons décrites plus haut, ou des techniques de recherche considérées de plus en plus comme « classiques » [Venturini, 2010]. Plus précisément, le terme fait référence à l'adaptation, inimaginable dans un autre contexte, de dispositifs et plates-formes réservés au Web (par exemple le moteur de recherche Google, Facebook et Wikipédia) à la recherche en sciences politiques et sociales. Parmi ces outils figure le développement du « Lippmannian Device », un Google Scraper

qui détecte les préférences ou les penchants d'un acteur en se fondant sur le type de mot-clé mentionné (voir Figure 6). On peut alors, en parcourant plusieurs sites Web sur les changements climatiques, effectuer une recherche par nom de climato-sceptiques et ainsi trouver des acteurs qui soient acquis à la cause (sans oublier les sites de veille qui les suivent ou les mentionnent). Dans ce cas, on a préféré réorienter Google en véritable module destiné à la recherche, plutôt que de l'utiliser, plus traditionnellement, en appareil conçu pour informer des clients

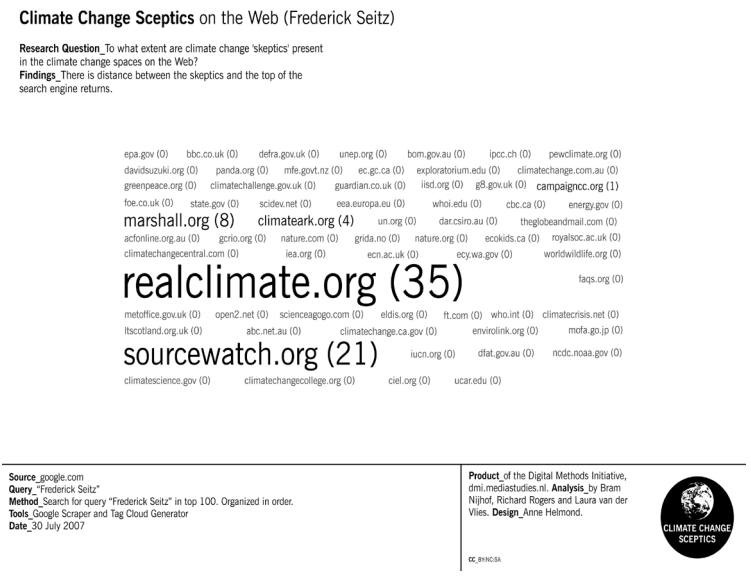

Climate Change Sceptics on the Web (Frederick Seitz)
Research Question_To what extent are climate change 'skeptics' present in the climate change spaces on the Web?
Findings_There is distance between the skeptics and the top of the search engine returns.
epa.gov (0) bbc.co.uk (0) defra.gov.uk (0) unep.org (0) bom.gov.au (0) ipcc.ch (0) pewclimate.org (0)
davidsuzuki.org (0) panda.org (0) mfe.govt.nz (0) ec.gc.ca (0) exploratorium.edu (0) climatechange.com.au (0)
greenpeace.org (0) climatechallenge.gov.uk (0) guardian.co.uk (0) iisd.org (0) g8.gov.uk (0) campaigncc.org (1)
foe.co.uk (0) state.gov (0) scidev.net (0) eea.europa.eu (0) whoi.edu (0) cbc.ca (0) energy.gov (0)
marshall.org (8) climateark.org (4) un.org (0) dar.csiro.au (0) theglobeandmail.com (0)
acfonline.org.au (0) gcrio.org (0) nature.com (0) grida.no (0) nature.org (0) ecokids.ca (0) royalsoc.ac.uk (0)
climatechangecentral.com (0) iea.org (0) ecn.ac.uk (0) ecy.wa.gov (0) worldwildlife.org (0)
realclimate.org (35) faqs.org (0)
metoffice.gov.uk (0) open2.net (0) scienceagogo.com (0) eldis.org (0) ft.com (0) who.int (0) climatecrisis.net (0)
ltscotland.org.uk (0) abc.net.au (0) climatechange.ca.gov (0) envirolink.org (0) mofa.go.jp (0)
sourcewatch.org (21) iucn.org (0) dfat.gov.au (0) ncdc.noaa.gov (0)
climatescience.gov (0) climatechangecollege.org (0) ciel.org (0) ucar.edu (0)
Source_google.com
Query_"Frederick Seitz"
Method_Search for query "Frederick Seitz" in top 100. Organized in order.
Tools_Google Scraper and Tag Cloud Generator
Date_30 July 2007
Product_of the Digital Methods Initiative, dmi.mediastudies.nl. **Analysis**_by Bram Nijhof, Richard Rogers and Laura van der Vlies. **Design**_Anne Helmond.
CC BY NC SA
CLIMATE CHANGE SCEPTICS

Figure 6 : Présence des climato-sceptiques sur les principaux sites consacrés aux changements climatiques selon http://google.com, juillet 2007. Source : analyse à distance par Google Scraper dit Lippmannian Device (dispositif Lippmanian), © Digital Methods Initiative, Amsterdam, 2007

Conclusion

On ne peut pas considérer les méthodes numériques, dans leur application générale ou dans celle, plus spécifique, de réorientation de dispositifs, comme de simples boîtes à outils ou comme le mode d'emploi d'une série de logiciels ; les méthodes numériques questionnent plus largement la manière de faire de la recherche en ligne. Elles invitent à considérer ou à imaginer d'un point de vue sociologique les opportunités de recherche qu'offre la culture en ligne, en se laissant guider par le médium au lieu de le faire répondre à des ordres disciplinaires, conceptuels. Prenons l'exemple, en guise de conclusion, de l'activisme politique. On pourrait observer d'un œil critique la montée du slacktivisme ou du clicktivisme, cette activité en ligne qui requiert (par un simple clic) un engagement minimum mais donne pour autant l'impression d'avoir été utile à une cause. On pourrait, à l'inverse, étudier dans quelle mesure aimer, partager et commenter un contenu est une preuve d'engagement, et par là-même étudier, par exemple, autour de quelles vidéos et quelles photos se rassemblent aujourd'hui les groupes et pages

anti-islamiques sur Facebook (voir Figure 7). L'étude de l'engagement s'inspire ici d'un cadre analytique qui capte aussi bien les clics que les commentaires et qui identifie le contenu qui rassemble, ouvrant ainsi la porte à d'autres interprétations. Il s'agit ici, dès la phase préliminaire, de s'appuyer sur les mesures d'activité du médium, et de se demander ensuite quels enseignements peuvent en être tirés.

Figure 7 : Contenus les plus suivis sur les réseaux Facebook européens anti-djihadistes, janvier 2013. Tiré de « What does the Internet add? Studying extremism and counter-jihadism online », International Workshop and Data Sprint. © Digital Methods Initiative, Amsterdam, 2013

Références bibliographiques

[Adobe, 2014] Adobe, 2014, « Adobe Analytics », http://www.adobe.com/solutions/digital-analytics.html.

[Ackland, 2013] Ackland, R., 2013, Web Social Science: Concepts, Data and Tools for Social Scientists in the Digital Age, London : Sage.

[Baetens *et al.*, 2013] Baetens, T., Juetten, T., Maessen, J., Borra, E. et Rogers, R., 2013, *De Uitzondering op de Regel: Over Ambtenaren in de Openbaarheid*, The Hague : Ministry of Internal Affairs/Emma Communicatie.

[Berry, 2011] Berry, D. M., 2011, « The Computational Turn: Thinking About the Digital Humanities », *Culture Machine*, 12, 1-22.

[Blood, 2002] Blood, R., 2002, « Introduction », In Rodzvilla J. (éd.), *We've Got Blog: How Weblogs Are Changing Our Culture*, Cambridge, MA : Perseus, ix-xiii.

[Bollena *et al.*, 2011] Bollena, J., Maoa, H. et Zengb, Z., 2011, « Twitter mood predicts the stock market », *Journal of Computational Science*, 2(1), 1-8.

Table des matières

Introduction 7

Marta Severo, Alberto Romele

Partie 1 - La trace, les méthodes et les données.....11

Au-delà de la critique *big data* 13

Richard Rogers

Les méthodes d'interface 33

Noortje Marres, Carolin Gerlitz

Soft data 63

Marta Severo, Alberto Romele

L'identité comme base de données 89

Jos de Mul

Partie 2 - Rencontre entre traces numériques et territoires 111

L'écume numérique des territoires 113

Dominique Boullier

Spatialités algorithmiques 135

Boris Beaude

L'hyperville 163

Franck Cormerais

Désirs de data 179

Maryse Carmes, Jean-Max Noyer

Partie 3 - Les pratiques de la trace numérique..... 213

De la trace à la carte et de la carte à la trace 215

Matthieu Noucher

Dimensions spatiales de l'actualité internationale..... 227

Laurent Beauguitte, Marta Severo

Reconfiguration des pratiques participatives 241

Nicolas Douay, Maryvonne Prévot

Remerciements 261

Les auteurs 263